Industrial Engineering
& Management Systems
Vol 22, No 4, December 2023, pp.389-398
ISSN 1598-7248 | EISSN 2234-6473 |

https://doi.org/10.7232/iems.2023.22.4.389
© 2023 KIIE

Enhancing Cascade Quality Prediction Method in Handling Imbalanced Dataset Using Synthetic Minority Over-Sampling Technique

Fajar Azhari Julian
Department of Industrial Engineering, Institut Teknologi Nasional Bandung, Indonesia

Fahmi Arif*
Department of Industrial Engineering, Institut Teknologi Nasional Bandung, Indonesia

(Received: September 13, 2022 / Revised: June 12, 2023 / Accepted: August 4, 2023)

ABSTRACT

Assessing the production process primarily revolves around quality. When dealing with a basic manufacturing process, quality can be easily anticipated. However, as manufacturing processes grow in complexity, it has been discovered through prior studies that directly predicting the quality of an intricate production system becomes challenging. This is due to the interdependency between each stage of manufacturing, where the outcomes of preceding stages impact subsequent processes. To address this issue, the Cascade Quality Prediction Method (CQPM) was developed. However, as this method employed a classification algorithm, CQPM cannot be directly applied when dealing with datasets that lack target variables for prediction or when the distribution of classes is imbalanced. This study aimed to improve the effectiveness of the CQPM in the context of multistage manufacturing. To achieve this, Hotelling's T2 and the Synthetic Minority Over-sampling Technique (SMOTE) algorithm were incorporated during the data preparation phase, especially when dealing with imbalanced class distributions and missing target variables. The results demonstrated that the inclusion of Hotelling's T2 allowed for the application of classification algorithms. By combining the CQPM approach with a random forest classifier and the SMOTE algorithm, notable improvements were observed in the model's performance. The enhanced data pre-processing techniques led to impressive metric values, including 99.95% accuracy, 93.75% G-Mean, and 96.76% F-Measure. These findings indicate the potential of the proposed model framework to accurately predict quality in multistage manufacturing systems.

Keywords: CQPM, Multistage Manufacturing, Quality Prediction, SMOTE, Hotelling's T2

* Corresponding Author, E-mail: fahmi.arif@itenas.ac.id

1. INTRODUCTION

Quality is something that is demanded by customers and is the main factor in assessing a production process (Kao *et al.*, 2017) and has become a major topic for many years (Durana *et al.*, 2019). To achieve these demands quality control is needed (Arif *et al.*, 2013a). Quality control seeks to monitor and predict the quality of products during the manufacturing processes. To predict the quality, a variable that becomes the target for the prediction is required from the dataset (Charbuty and Abdulazeez, 2021). In a simple manufacturing process, the target variable is available. So the quality can be predicted straightforwardly (Arif *et al.*, 2013b).

Under certain conditions, a manufacturing dataset has a lot of error value or is not detected in the manufacturing production process dataset, causing the data cannot be processed directly (Oleghe, 2020) and because of its nature, the manufacturing process inevitably generates an unbalanced dataset (Cheng *et al.*, 2018) as it happened in the multi-stage continuous-flow manufacturing process which is reflected in the dataset from Liveline Technologies (Supergus, 2019). There is in which the number of defective products is smaller (minority) compared to good products (majority), this is called imbalanced and will affect the predicted value (Leevy *et al.*, 2018).

Furthermore, the complexity of manufacturing processes caused the dataset to become more complicated with high-dimensional variables, containing large amounts of data with various input and output variables, and uncertain and dynamic environments (Cheng *et al.*, 2018). This complex manufacturing process is known as Multi-stage Manufacturing System (MMS). Where each stage will be influenced by the previous stage, and affect the next process so that the relationship between the production process is more complex (Lee *et al.*, 2020).

Cascade Quality Prediction Method from Arif *et al.*, (2013a) is employed. Quality prediction in previous research (Ismail *et al.*, 2021) use unsupervised and supervised machine learning methods, where the unsupervised use PCA and K-Means algorithm and the supervised use SVM, ANN, KNN, and RF, and the result show that the supervised with CQPM framework is better with the average performance of sensitivity and specificity is increased by 4.76% and 3.41%. Arif *et al.* (2013b) used the CQPM in their research to partially predict the quality, ID3, and PCA on the semiconductor dataset, the result is the ID3 and PCA method generates the highest accuracy value (90%) with a low G- mean Value (44%). However, in this research, the dataset used has a target variable and does not pay attention to the imbalanced dataset. Therefore, the research (Chazhoor *et al.*, 2020) proves that using SMOTE on the semiconductor manufacturing (SECOM) data set can handle the imbalanced dataset and improve prediction accuracy.

This research proposes the CQPM with the addition of a multivariate statistical process and resampling method. Multivariate statistics such as Hotelling's T2 method are intended to obtain one target variable from data with multivariate variables (Murphy, 1987). SMOTE is employed to overcome the imbalanced data caused by erroneous (Fernández *et al.*, 2018). and machine learning techniques such as Decision Tree, Random Forest, Naïve Bayes, k-Nearest Neighbor, and Support Vector Machine are used to support the CQPM framework to predict and extract insight from multistage manufacturing datasets (Wuest *et al.*, 2016).

Some researchers used machine learning and data mining in their research to handle the classification process such as quality prediction in MMS systems, such as (Kao *et al.*, 2017) by developing the Naive-Bayes and PCA on the semiconductor dataset. A Random Forest classifier model was used to evaluate and predict the feature selection subset (Oleghe, 2020). Research (Peres *et al.*, 2019) shows that non-linear models like XGBoost and Random Forests can model the complexity of such an environment, achieving a high true positive rate and showing promise for the improvement of existing quality control approaches, enabling defects and deviations to be addressed earlier and thus assist in reducing scrap and repair costs.

2. MATERIALS AND METHODOLOGY

Performing a data mining project requires guidance as a standard approach that will help translate business problems into data mining tasks, suggest appropriate data transformations and data mining techniques, and provide means for evaluating the effectiveness of the results and documenting the experience. The Cross Industry Standard Process for Data Mining (CRISP-DM) methodology is a comprehensive process model that is described in six phases which are: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Wirth and Hipp, 2000).

2.1 Business Understanding

The datasets provided by Liveline Technologies on Kaggle (Supergus, 2019) offer valuable insights as they are derived from real multi-stage manufacturing system processes. These datasets encompass a wide range of production processes, including those involving parallel activities as well as serial production processes at the assembly point. With a ti me dimension, the datasets capture different types of data collected from multiple workstations, enabling the analysis of temporal aspects and patterns within the manufacturing system. By mining these datasets, researchers and analysts can uncover significant information and gain a deeper understanding of the intricacies of multi-stage manufacturing processes.

2.2 Data Understanding

According to the dataset providers, the raw dataset consists of 14,088 observations with a time stamp at a 1-second interval. The dataset contains 116 variables. The key variables of interest, as depicted in Figure 1, are the measurements of 15 stages 1 outputs and 15 stages 2 outputs. These measurements are expressed in millimeters, indicating that the dataset primarily consists of dimensional features. Notably, this dataset does not include any target variables or labels.

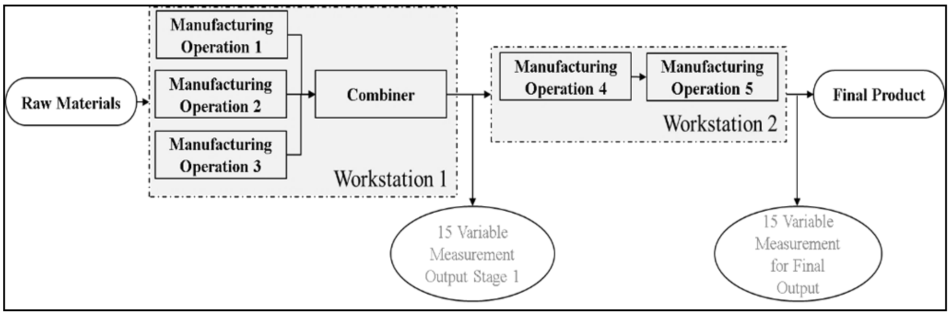
Raw Materials
Manufacturing Operation 1
Manufacturing Operation 2
Manufacturing Operation 3
Combiner
Workstation 1
Manufacturing Operation 4
Manufacturing Operation 5
Workstation 2
Final Product
15 Variable Measurement Output Stage 1
15 Variable Measurement for Final Output

Figure 1. Flow process multi-stage manufacturing system.

2.3 Data Preparation

Because of the complexity of the dataset, it is necessary to do data preparation to maximize the results of the model and the classification made. Following are some steps in data preparation.

Step 1: Data cleansing

The dataset used has a lot of error (noise) values, hence, denoising needs to be done. The study (Hildebrand and Mathsson, 2020) determined a percentage of 10-20% of the error value as the limit in the selected variable to be deleted. In research (Oleghe, 2020) the selection of variables was done by deleting variables that had a 70% error value, but this cannot be applied to this study because it will throw away many necessary variables. Determining the error value limit in a dataset is uncertain, each dataset has its own uniqueness (Oleghe, 2020). Therefore, in this study, the error value limit used is 25%, if a variable has an error value of 25%, it will be deleted. Elimination of variables that have error values exceeding 25% is removed because research (Bohannon *et al.*, 2005) shows that replacing the value of the error with a prediction using a classification model will result in excessive costs and computational time. This was justified by Oleghe (2020) because to make a prediction, it will require a different variety of model classifications to correct for the error value of the variable, considering that this research is not focused only on denoising, the variables and error values are deleted.

Step 2: Data standardization

The data consist of various dimensions and units, this can hinder the prediction process because the data is not well structured (Skarpathiotaki and Psannis, 2022). Data standardization is done to change data with dynamic range values to be more specific and structured (Mohamad and Usman, 2013).

Step 3: Data transformation

As previously mentioned in the data understanding section, the absence of a target variable in the dataset makes it unsuitable for applying classification algorithms. However, Figure 1 depicts measurements taken after each production stage, specifically capturing the output related to quality variables. Since there are multiple variables involved in these measurements, the use of multivariate control charts becomes relevant for examining the process state and determining whether it is in control or out of control. In this scenario, Hotelling's T2, a statistical technique, is employed. Hotelling's T2 utilizes multivariable quality control panels, enabling the monitoring of all variables using a single panel. This approach proves particularly valuable when there are interrelationships among the variables. The fundamental principle behind these panels involves quantifying the deviations of all variables from their average value at a specific moment while considering the differences between these values and the mean. As a result, a new variable is generated that represents the target value, indicating whether the process is in-control or out-of-control state.

Step 4: Resampling

After performing a data transformation step, the resulting dataset now includes both predictor and target variables. However, due to the nature of the manufacturing process, the distribution of the in-control and out-of-control classes is imbalanced. To overcome this issue, resampling methods are commonly employed in the data preprocessing phase. One prevalent technique in this regard is SMOTE, which addresses the imbalance by adjusting the sizes of both the minority and majority classes, resulting in a balanced distribution within the training dataset. This approach proves effective in mitigating the challenges associated with imbalanced datasets, as supported by previous research (Galar *et al.*, 2012) and (Kang *et al.*, 2017).

Step 5: Splitting dataset into train-test data

To enable the validation process, it is essential to partition the dataset into two distinct parts: the training dataset and the test dataset. This train/test split is a reliable and straightforward validation approach where the dataset is divided before the mining or modeling phase, and it is used only once to validate the developed model. Previous studies by Amari *et al.* (1997) and Vabalas *et al.*

(2019) support the practice of allocating 80% of the dataset to the training dataset and 20% to the test dataset. This partitioning helps mitigate problems such as overfitting and overtraining. The study demonstrates that by adopting an 80:20 split, the model can effectively adapt to the data and achieve high accuracy.

2.4 Modeling

In a multistage manufacturing system quality prediction model has two different approaches, single and multi-point approaches. A single-point approach is a holistic quality prediction model used to predict the whole manufacturing process as illustrated in Figure 2.

From the illustration above, the value of the quality prediction is obtained in the final workstation regardless of the connection between workstations. Therefore, this approach can explain the correlation between manufacturing operations from one workstation to another. In this condition, the final quality can be expressed by:

$$Q = f\,(Ws) \tag{1}$$

where:

Q= quality prediction
Ws = manufacturing operation variable in the workstation.

Otherwise, a multi-point approach is developed to explain the correlation between the workstations as illustrated in Figure 3.

However, this approach will cause redundancy because it is done repeatedly on every workstation. In this approach, the final quality can be expressed by:

$$Q_n = f\,(X_n) \tag{2}$$

where:

Q_n = final product quality
X_n = quality prediction in n^{th} workstation, n=1, 2, 3, …

By considering the partial and final quality prediction in a workstation as shown in Figure 3 and Figure 4 a new approach was developed by Arif *et al.* (2013a) called CQPM. The relationship among variables in the CQPM method can be illustrated as shown in Figure 4.

For this condition, the relationship between variables is MMS can be expressed by:

$$y_i, k = f\,(\,y_{i-1,\,k}, X_{i,j}) \tag{3}$$

$$Q = f\,(\,y_{n,k}) \tag{4}$$

where:

$X_{i,j} = j^{\text{th}}$ manufacturing operation variable in i^{th} workstation, i = 1,2,3, … n and j = 1,2,3, … m
$y_{i,k} = k^{\text{th}}$ characteristics of output from an i^{th} workstation, k = 1,2,3…
Q= final product quality

Equation (3) establishes a relationship between workstations by utilizing the single-point approach and multi-point approach, as explained earlier. The output characteristics from a particular workstation are influenced not only by the manufacturing operation within that workstation but also by all preceding workstations. The relationship between the operation process and the final product is observed in the final workstation, denoted as $y_{n,k}$. Additionally, the quality of the product (Q) is

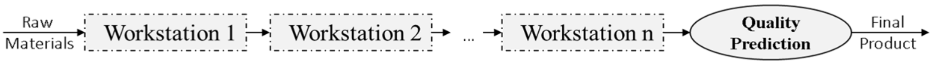

Figure 2. Quality prediction MMS single-point approach.

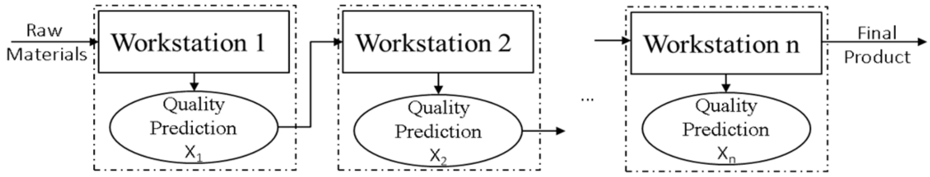

Figure 3. Quality Prediction MMS Multi-Point Approach.

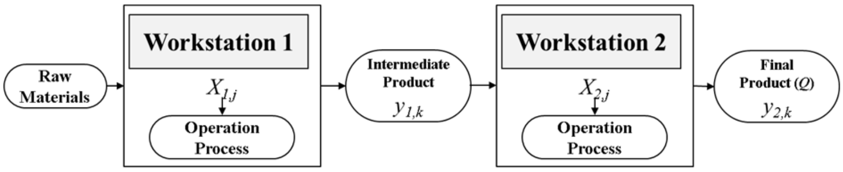

Figure 4. The relationship variable between operations in each workstation MMS.

represented in equation (4). Typically, the product quality is determined based on whether it is accepted or rejected. However, in this study, the data collected comprises multivariate variables. Therefore, to address equation (4), the Hotelling's T2 method is initially applied to process the data. Afterward, the problem can be approached as a classification problem.

2.5 Prediction Model

The manufacturing line under study was divided into two stages, as depicted in Figure 1. In Stage 1, three machines are operating in parallel. Once a product is completed by a machine, it proceeds to the next stage where it combines with products from other machines. The output is measured at 15 separate locations. In Stage 2, the output from Stage 1 is conveyed to the second stage through a conveyor. Stage 2 comprises two machines operating in series. After the product completes its processing in the fifth machine, its output is measured at 15 distinct locations.

The framework concept is comprised of multiple models, beginning with Model A, which involves two prediction points. This model adopts a multi-point approach and is designed to predict the output of each stage independently. In the prediction model for Stage 1, referred to as Model A1, ambient conditions and parameters from machines 1-3 are considered, as depicted in Figure 1. Conversely, the prediction model for Stage 2, or Model A2, focuses on ambient conditions and parameters from machines 4-5, as illustrated in Figure 6.

The second prediction model, Model B, adopts a multi-stage manufacturing approach with a single-point prediction. In this model, the process flows directly from Stage 1 to Stage 2 without considering the output from Stage 1. Instead, the final output is determined solely by the output from Stage 2. The single-stage approach in

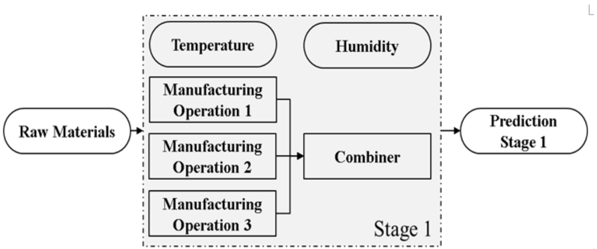
Temperature
Humidity
Raw Materials
Manufacturing Operation 1
Manufacturing Operation 2
Manufacturing Operation 3
Combiner
Prediction Stage 1
Stage 1

Figure 5. Model A_1 multi-stage multi-point approach quality prediction stage 1.

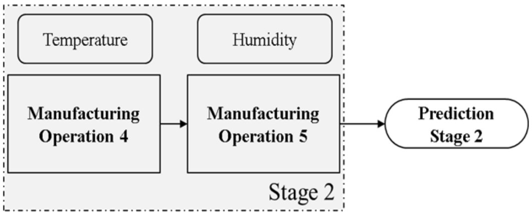

Figure 6. Model A_2 multi-stage multi-point approach quality prediction stage 2.

Model B involves using specific parameters from each machine and the ambient condition, as depicted in Figure 7. These parameters are utilized to predict the final output of the manufacturing process.

The final prediction model, Model C, adapts the CQPM framework. In this model, the output from Stage 1 is utilized as input for Stage 2. Hence, Model C considers the stage 1 output, along with ambient conditions and parameters from machines 4-5, as predictor variables for the model. Figure 8 illustrates this model, highlighting the relationship between these variables within the CQPM framework.

To investigate the relationships between variables, the three specified models (Model A, Model B, and Model C) underwent an exploration using a diverse range of

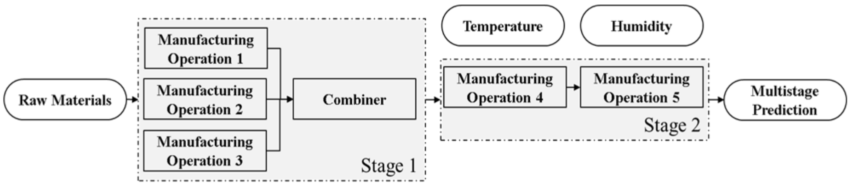
Temperature
Humidity
Raw Materials
Manufacturing Operation 1
Manufacturing Operation 2
Manufacturing Operation 3
Combiner
Stage 1
Manufacturing Operation 4
Manufacturing Operation 5
Stage 2
Multistage Prediction

Figure 7. Model B multi-stage single-point approach quality prediction.

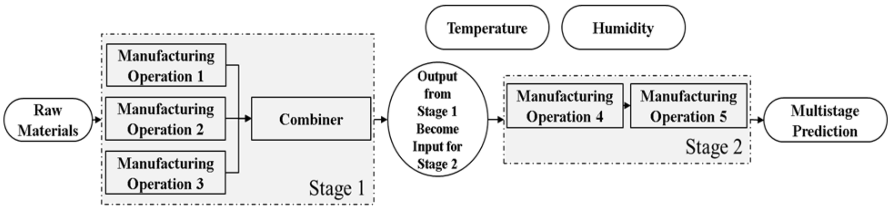
Temperature
Humidity
Raw Materials
Manufacturing Operation 1
Manufacturing Operation 2
Manufacturing Operation 3
Combiner
Stage 1
Output from Stage 1 Become Input for Stage 2
Manufacturing Operation 4
Manufacturing Operation 5
Stage 2
Multistage Prediction

Figure 8. Model C multi-stage CQPM approach quality prediction.

classification algorithms. These algorithms aim to construct classification models based on dataset characteristics and classify unknown data objects into known categories. Commonly utilized algorithms suitable for datasets with nominal data types include decision trees (DT), random forests (RF), naïve Bayes (NB), k-nearest neighbor (KNN), and support vector classifiers (SVC). Each algorithm was sequentially applied to each model, resulting in the creation of multiple prediction models. Subsequently, the performance of these models was assessed to determine the top-performing model among Model A, Model B, and Model C. To facilitate a comprehensive comparison, these algorithms were also applied to both balanced and imbalanced datasets.

2.6 Evaluation

After completing the experiment, a test was conducted to verify if the datasets met the specified requirements for prediction. The detailed results and discussions can be found in the corresponding section. If the obtained results aligned with the predetermined expectations, the product was categorized as accepted; otherwise, it was deemed rejected. Performance measurement plays a crucial role in the evaluation of classifications. The confusion matrix stands as one of the most employed tools for assessing classification performance. The confusion matrix consists of four classes: TP (True Positive) represents the number of correctly classified positive examples, TN (True Negative) represents the number of correctly classified negative examples, FP (False Positive) represents the number of negative examples incorrectly classified as positive, and FN (False Negative) represents the number of positive examples misclassified as negative (Chawla, 2010). By utilizing the information provided by the confusion matrix, performance criteria such as accuracy, G-Mean, and F-Measure can be determined (Tahir *et al.*, 2019).

Accuracy (Acc) is a widely employed evaluation metric that assesses the correctness of a model's predictions. It is commonly presented as a percentage and indicates the proportion of correctly classified instances out of the total number of instances in the dataset. The calculation of accuracy can be represented by Equation (5).

$$\mathrm{Acc} = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

G-mean, also known as the geometric mean or balanced accuracy, is a commonly utilized evaluation metric in machine learning, particularly in the presence of imbalanced datasets. It combines sensitivity (true positive rate) and specificity (true negative rate) to offer a balanced assessment of a model's performance. The objective of G-mean is to maximize the classification accuracy for the entire population by achieving a balance between the accuracy of positive and negative instances (Yu *et al.*, 2012). Equation (6) illustrates the calculation of G-mean.

$$G - Mean = \sqrt{\frac{TN}{TN + FP} x \frac{TP}{TP + FN}} \tag{6}$$

F-measure, also referred to as F1 score, is a widely employed evaluation metric in machine learning, primarily used for assessing the performance of binary classification models. By combining precision and recall, it offers a balanced evaluation of the model's effectiveness. The F-measure proves particularly valuable in scenarios where there is an imbalance between the number of positive and negative instances in the dataset (Tahir *et al.*, 2019). It provides a comprehensive assessment by considering both the model's capability to correctly identify positive instances and its ability to minimize false positives.

The F-measure calculates the harmonic mean of precision and recall, resulting in a single metric that strikes a balance between the two. This ensures that the evaluation metric considers both the precision (accuracy of positive predictions) and recall (ability to capture positive instances). The F-measure is expressed mathematically as shown in Equation (7).

$$F - Measure = 2\,x\,\frac{\frac{TP}{TP + FP} x \frac{TP}{TP + FN}}{\frac{TP}{TP + FP} + \frac{TP}{TP + FN}} \tag{7}$$

where:

TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative

3. RESULT AND DISCUSSION

In this study, this mining process is implemented to the continuous_factory_process.csv dataset using Python 3.0 and Ms. Excel on a 2.40 GHz computer with 4.00 GB memory. The most important aspect that quality monitoring goes well is to be able to detect rejected or defective products with a minimum rate of error. These two factors can be described by G-mean and F-measure metrics. The accuracy in this study is remarkably high because five machine learning algorithms work well in every model to predict product quality.

Table 1 and Table 2 present the performance evaluation of the quality prediction model using a multi-stage multi-point approach, as depicted in Figure 5 and Figure 6. The tables showcase the results obtained when implementing classification algorithms on imbalanced datasets. It is observed that this approach initially fails to generate a satisfactory prediction model, as evidenced by the low

F-Measure values. Some algorithms even exhibit errors when attempting to predict the True Negative class.

However, after applying the SMOTE algorithm, significant improvements are observed in the results. The model generated by the K-Nearest Neighbor algorithm achieves the best performance, exhibiting an Accuracy of 99.47%, a G-Mean of 99.73%, and an F-Measure of 83.78%. These metrics indicate a substantial enhancement in the model's predictive capabilities, demonstrating the effectiveness of the SMOTE algorithm in addressing the challenges posed by imbalanced datasets.

Table 3 presents the performance evaluation of the model when the multi-stage single point approach is applied. Similar to the previous approach, the results indicate that implementing this approach on the imbalanced dataset does not yield a high-quality prediction model. The performance metrics, particularly the F-Measure, suggest suboptimal results. However, the introduction of the SMOTE algorithm to balance the dataset proves to be effective in improving the model's performance. The application of the SMOTE algorithm enhances the model's ability to handle the class imbalance issue, leading to more reliable predictions. Among the models evaluated, the Random Forest model achieves the best performance, exhibiting an Accuracy of 97.52%, a G-Mean of 99.76%, and an F-Measure of 76.85%. These metrics demonstrate notable improvements in the model's predictive capabilities after employing the SMOTE algorithm, resulting in a more balanced and accurate prediction model.

Table 4 displays the performance evaluation of the model when the CQPM approach is applied. The results indicate that implementing this approach yields superior performance compared to the other approaches evaluated. This finding aligns with the research conducted by Arif *et al.* (2013b) and Ismail *et al.* (2021), which highlight that considering the input-output relationship of each stage within the CQPM model leads to improved quality prediction values compared to other models.

Table 1. Performance of classifier algorithm in prediction model A_1

		Imbalanced Dataset					Synthetic Minority Oversampling Technique				
	Classifier	DT	RF	NB	KNN	SVC	DT	RF	NB	KNN	SVC
Metric	Accuracy	99.84%	99.79%	86.28%	99.89%	99.79%	99.47%	99.79%	71.03%	99.63%	92.64%
	G-Mean	62.50%	50.00%	68.18%	75.00%	50.00%	99.73%	99.89%	60.54%	99.81%	96.31%
	F-Measure	76.90%	NaN	57.91%	85.69%	NaN	78.18%	85.68%	54.83%	81.02%	67.03%
Confusion Matrix	TP	1884	1884	1627	1884	1884	1874	1880	1339	1877	1745
	FP	0	0	257	0	0	10	4	545	7	139
	FN	3	4	2	2	4	0	0	2	0	0
	TN	1	0	2	2	0	4	4	2	4	4

Table 2. Performance of classifier algorithm in prediction model A_2

		Imbalanced Dataset					Synthetic Minority Oversampling Technique				
	Classifier	DT	RF	NB	KNN	SVC	DT	RF	NB	KNN	SVC
Metric	Accuracy	99.84%	99.31%	95.50%	99.84%	99.84%	96.50%	99.74%	94.49%	99.42%	81.14%
	G-Mean	50.00%	49.73%	47.82%	50.00%	50.00%	81.61%	99.87%	47.32%	99.71%	73.92%
	F-Measure	NaN	49.83%	48.85%	NaN	NaN	63.12%	81.44%	48.58%	75.47%	59.83%
Confusion Matrix	TP	1885	1875	1803	1885	1885	1820	1880	1784	1874	1530
	FP	0	10	82	0	0	65	5	101	11	355
	FN	3	3	3	3	3	1	0	3	0	1
	TN	0	0	0	0	0	2	3	0	3	2

Table 3. Performance of classifier algorithm in prediction model B

		Imbalanced Dataset					Synthetic Minority Oversampling Technique				
	Classifier	DT	RF	NB	KNN	SVC	DT	RF	NB	KNN	SVC
Metric	Accuracy	99.63%	99.79%	93.01%	99.89%	99.84%	97.19%	99.52%	92.85%	99.36%	82.26%
	G-Mean	66.53%	49.97%	46.58%	66.67%	50.00%	81.95%	99.76%	46.50%	99.68%	91.11%
	F-Measure	62.13%	49.95%	48.19%	79.98%	NaN	63.50%	76.85%	48.15%	74.91%	64.94%
Confusion Matrix	TP	1880	1884	1756	1885	1885	1833	1876	1753	1873	1550
	FP	5	1	129	0	0	52	9	132	12	335
	FN	2	3	3	2	3	1	0	3	0	0
	TN	1	0	0	1	0	2	3	0	3	3

Table 4. Performance of classifier algorithm in prediction model C

		Imbalanced Dataset					Synthetic Minority Oversampling Technique				
	Classifier	DT	RF	NB	KNN	SVC	DT	RF	NB	KNN	SVC
Metric	Accuracy	99.58%	99.63%	96.08%	99.63%	99.68%	99.05%	99.95%	94.39%	99.47%	93.70%
	G-Mean	74.89%	62.47%	73.14%	68.70%	62.50%	99.52%	93.75%	72.29%	99.73%	96.84%
	F-Measure	74.89%	71.35%	61.19%	73.86%	76.88%	78.92%	96.76%	60.34%	83.78%	68.63%
Confusion Matrix	TP	1876	1879	1810	1878	1880	1862	1880	1778	1870	1761
	FP	4	1	70	2	0	18	0	102	10	119
	FN	4	6	4	5	6	0	1	4	0	0
	TN	4	2	4	3	2	8	7	4	8	8

The present study further supports these findings by demonstrating that the predictive value of the CQPM model, when applied to imbalanced data, outperforms the prediction models produced by the multi-stage single point and multi-stage multi-point approaches. Moreover, the CQPM model shows better performance without errors and exhibits high predictive value. These findings reinforce the effectiveness of the CQPM approach in accurately predicting the quality of the production process, making it a favorable choice compared to the other evaluated approaches.

While the CQPM approach demonstrates better performance compared to other approaches, it is important to note that the G-Mean and F-Measure values obtained are still relatively low. This suggests an imbalance in the model's ability to predict both the majority and minority classes accurately. It is crucial to consider that a high accuracy metric value does not guarantee satisfactory results, as emphasized in studies by Chawla (2010) and Tahir *et al.* (2019). This study further confirms that high accuracy values do not always correspond to high G-Mean and F-Measure metric values.

Research by Sáez *et al.* (2015) establishes that resampling techniques can be employed to enhance the predictive value of quality in imbalanced datasets. In line with this, the current study demonstrates that implementing the SMOTE algorithm leads to an improvement in prediction performance. Processing the imbalanced data using the CQPM model with the random forest classification algorithm yields an accuracy value of 99.63%, a G-Mean of 62.47%, and an F-Measure of 71.35%. However, employing the SMOTE resampling algorithm with the same model and classification algorithm significantly enhances the results, resulting in an accuracy value of 99.95%, a G-Mean of 93.75%, and an F-Measure of 96.76%. This difference highlights the substantial impact of the SMOTE algorithm on improving the predictive value of the model, particularly in terms of achieving a balanced performance across different evaluation metrics.

4. CONCLUSION AND FUTURE WORKS

Hotelling's T2 method enables the creation of new classification variables from multivariate variables. In the context of multi-stage manufacturing systems, the CQPM method has been identified as the most effective approach. However, it fails to address the challenges posed by imbalanced data. Therefore, to handle imbalanced datasets, the recommended solution is to employ the SMOTE technique. Additionally, the combination of the CQPM approach with the random forest classifier algorithm proves to be the best classifier among all the models considered, achieving a high predictive value in the dataset with 99.95% accuracy, 93.75% G-mean, and 96.76% F-measure.

Future work should focus on further improving the model performance by predicting error values using machine learning techniques, adopting advanced feature selection methods, exploring alternative multivariate statistical control approaches, and exploring different classification techniques. These efforts aim to enhance the predictive capabilities of the model and achieve better results in future predictions.

ACKNOWLEDGMENT

This work has been financially supported by the Faculty of Industrial Technology, Institut Teknologi Nasional Bandung, Indonesia. A great appreciation to the faculty members for their support in this project.

REFERENCES

Arif, F., Suryana, N., and Hussin, B. (2013a), Cascade quality prediction method using multiple PCA+ID3 for multi-stage manufacturing system, *IERI Procedia*, **4**, 201-207.

Amari, S. I., Murata, N., Muller, K. R., Finke, M., and Yang, H. H. (1997), Asymptotic statistical theory of overtraining and cross-validation, *IEEE Transactions on Neural Networks*, **8**(5), 985-996.

Arif, F., Suryana, N., and Hussin, B. (2013b), A data mining approach for developing quality prediction model in multi-stage manufacturing, *International Journal of Computer Applications*, **69**(22), 35-40.

Bohannon, P., Fan, W., Flaster, M., and Rastogi, R. (2005), A cost-based model and effective heuristic for repairing constraints by value modification, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 143-154.

Charbuty, B. and Abdulazeez, A. (2021), Classification based on decision tree algorithm for machine learning, *Journal of Applied Science and Technology Trends*, **2**(1), 20-28.

Chawla, N. V. (2010), Data mining and knowledge discovery handbook, *Data Mining and Knowledge Discovery Handbook*, 2003-2004, Available from: https://doi.org/10.1007/978-0-387-09823-4.

Chazhoor, A., Mounika, Y., Vergin Raja Sarobin, M., Sanjana, M. V., and Yasashvini, R. (2020), Predictive maintenance using machine learning based classification models, *IOP Conference Series: Materials Science and Engineering*, **954**(1), 012001.

Cheng, Y., Chen, K., Sun, H., Zhang, Y., and Tao, F. (2018), Data and knowledge mining with big data towards smart production, *Journal of Industrial Information Integration*, **9**, 1-13.

Durana, P., Kral, P., Stehel, V., Lazaroiu, G., and Sroka, W. (2019), Quality culture of manufacturing enterprises: A possible way to adaptation to industry 4.0, *Social Sciences*, **8**(4), 124.

Fernández, A., Garcia, S., Herrera, F., and Chawla, N. V. (2018), SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary, *Journal of Artificial Intelligence Research*, **61**, 863-905.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., and Herrera, F. (2012), A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches, *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, **42**(4), 463-484.

Hildebrand, C. and Mathsson, M. (2020), *Removing noise with an Autoencoder in a Predator-Prey Ordinary Differential Equation*, KTH Royal Institute of Technology.

Ismail, M., El-Assal, A., and Mostafa, N. A. (2021), Utilization of machine learning techniques for quality monitoring and prediction, *Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management*, Singapore, 4830-4839.

Kang, Q., Chen, X. S., Li, S. S., and Zhou, M. C. (2017), A noise-filtered under-sampling scheme for imbalanced classification, *IEEE Transactions on Cybernetics*, **47**(12), 4263-4274.

Kao, H. A., Hsieh, Y. S., Chen, C. H., and Lee, J. (2017), Quality prediction modeling for multistage manufacturing based on classification and association rule mining, *MATEC Web of Conferences*, 123.

Lee, D. H., Yang, J. K., Kim, S. H., and Kim, K. J. (2020), Optimizing mean and variance of multiresponse in a multistage manufacturing process using operational data, *Quality Engineering*, **32**(4), 627-642

Leevy, J. L., Khoshgoftaar, T. M., and Bauder, R. A. and Seliya, N. (2018), A survey on addressing high-class imbalance in big data, *Journal of Big Data*, **5**(1), 1.

Mohamad, I. B. and Usman, D. (2013), Standardization and its effects on K-means clustering algorithm, *Research Journal of Applied Sciences, Engineering, and Technology*, **6**(17), 3299-3303, Available from: https://doi.org/10.19026/rjaset.6.3638.

Murphy, B. J. (1987), Selecting out of control variables with the T^2 multivariate quality control procedure, *The Statistician*, **36**(5), 571-581.

Oleghe, O. (2020), A predictive noise correction methodology for manufacturing process datasets, *Journal of Big Data*, **7**(1).

Peres, R. S., Barata, J., Leitao, P., and Garcia, G. (2019), Multistage quality control using machine learning in the automotive industry, *IEEE Access*, **7**, 79908-79916.

Sáez, J. A., Luengo, J., Stefanowski, J., and Herrera, F. (2015), SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering, *Information Sciences*, **291**, 184-203.

Skarpathiotaki, C. G. and Psannis, K. E. (2022), Cross-industry process standardization for text analytics, *Big Data Research*, **27**, 100274.

Supergus (2019), Multi-stage continuous-flow manufacturing process, cited 2020 March 20, Available from: https://www.kaggle.com/datasets/supergus/multistage-continuousflow-manufacturing-process.

Tahir, M. A. U. H., Asghar, S., Manzoor, A., and Noor, M. A. (2019), A classification model for class imbalance dataset using genetic programming, *IEEE Access*, **7**, 71013-71037.

Vabalas, A., Gowen, E., Poliakoff, E., and Casson, A. J. (2019), Machine learning algorithm validation with a limited sample size, *PLoS ONE*, **14**(11), e0224365.

Wirth, R. and Hipp, J. (2000), CRISP-DM: Towards a standard process model for data mining, *Proceedings of the 4th International Conference on the Practical Application of Knowledge Discovery and Data Mining*, 29-39.

Wuest, T., Weimer, D., Irgens, C., and Thoben, K. D. (2016), Machine learning in manufacturing: Advantages, challenges, and applications, *Production and*

Manufacturing Research, **4**(1), 23-45.

Yu, H., Ni, J., Dan, Y., and Xu, S. (2012), Mining and integrating reliable decision rules for imbalanced cancer gene expression data sets, *Tsinghua Science and Technology*, **17**(6), 666-673.

Fajar Azhari Julian achieved his master's degree in industrial engineering from Institut Teknologi Nasional Bandung in 2022. Currently, he is employed in a private company, where data science constitutes a fundamental aspect of his daily responsibilities. Simultaneously, he remains an ardent independent researcher, nurturing his passion for ongoing exploration in this domain.

Fahmi Arif holds a Ph.D. in Industrial Computing. In addition to his formal education, he has achieved several professional certifications, including Data Science Professional. His primary research interests encompass data science, artificial intelligence, and machine learning, especially their practical applications in industrial automation.